POTE Egészségügyi Szervezési Intézet

A kórházi adatfeldolgozás néhány matematikai, illetve számítógépes problémája

Gaál Aladár

Az egészségügy egyik legsürgetőbb, megoldásra váró feladata a mindmáig korszerűtlen kórházi, rendelőintézeti adattárolás átszervezése, computerizálása. A hagyományos kartotékrendszert felszámolva, meg kell honosítanunk az elektromágneses jelrögzítést. Ez nem csupán az adatok gyorsabb hozzáférhetősége, s így a gyógyítás hatékonyságának növelése érdekében indokolt, hanem a további tudományos kutatómunka szempontjából is elengedhetetlen. Ugyanis a biológiai ismeretek komoly hányada empirikus - így a kórházak, klinikák, laboratóriumok óriási adatmennyiségéből leszűrhető tapasztalatok az alap- és alkalmazott kutatások irányvonalait is kijelölhetik.

Vitathatatlan, hogy az empirikus összefüggések alátámasztásához megkívánt nagy számu kisérlet, mérési eredmény korszerű, gyors és megbízható kiértékeléséhez nélkülözhetetlen a computerek felhasználása. Tehát a számítógépek nemcsak az adattárolás racionalizálása érdekében, hanem a bennük tárolt információk, adatok közti összefüggések felismerésének megkönnyítéséhez is szükségesek. Ezért kell nagy sulyt fektetnünk a számítógépes adatfeldolgozó módszerek kifejlesztésére.

Első lépésként az alapvető metodikai elveket, problémákat kell tisztáznunk.

A vizsgált adathalmaz változói között fennálló kapcsolatok elemzésekor két lehetőség áll előttünk:

A.) A biológiai folyamat részletes analizisével a vizsgált jelenséget több apró részfolyamatra bontjuk, s megpróbáljuk ezek számszerű kapcsolatát megállapitani. Tehát megalkotjuk a jelenség modelljét, melyben a biológiai, kémiai, fizikai, stb. hatásokat - az alkalmasan választott matematikai apparátussal mintegy összefogjuk. A szakismeretek mélységétől függően változó tényezőt is be kell iktatnunk a modellbe. E modellt - valószinüségi változóiról - sztochasztikusnak nevezzük.

B.) Van azonban egy másik lehetőség is. Nem keressük a folyamat belső, lényegi összefüggéseit - nem a "miértekre", hanem

csak a "hogyanokra" akarunk választ kapni.

Elegendő számunkra a folyamat kezdeti (ill. bemenő) és vég (ill. kimenő) paraméterei között tapasztalható közvetlen, akár közelítő összefüggés ismerete is. Ugyanis sok esetben lehetetlen - a biológiai folyamatok bonyolultsága miatt - az ok-okozati összefüggéseket végigkövetni. A kauzális helyett sztochasztikus szkémákkal kell dolgoznunk. Igy többnyire meg kell elégednünk a jelenségre jellemző alapvető tendenciák, trendvonalak ismeretével.

I. Minőségi ismérvek:

Ha kvalitativ változókkal van dolgunk, ugy ezek előfordulásának relativ gyakoriságából tudunk következtetéseket levonni. Tekintve, hogy a vizsgált esetek általában összetettek, igy nem is az egyes ismérvek, hanem ezek különböző kombinációinak gyakoriságára vagyunk kiváncsiak.

Tehát kigyüjtjük az adathalmazból, hogy hányszor fordultak elő pl. az alábbi tulajdonságok együtt:

$X_{11} \wedge X_{21}$

$X_{31} \wedge X_{21}$

$X_{41} \wedge X_{21}$

. .
. .
. .

$X_{n1} \wedge X_{21}$

Ezen belül pl. $X_{11} \wedge X_{22}$; $X_{12} \wedge X_{22}$;

$X_{12} \wedge X_{21}$; $X_{11} \wedge X_{21}$ stb, stb. $X_{1m} \wedge X_{2p}$

Hasonlóképpen kigyüjthetők a három, négy, öt stb. változókat tartalmazó csoportok is:

pl. $X_{11} \wedge X_{21} \wedge X_{31}$ stb.

$X_{11} \wedge X_{22} \wedge X_{35} \wedge X_{49} \wedge \ldots \wedge X_{nm}$ stb, stb.

(Megjegyzés: az első index a tulajdonságra, a második annak állapotára utal.)

Kombinatórikai ismeretek birtokában belátható, hogy - a változók különböző állapotából adódó eshetőségeket nem is tekintve, - az egyes változó-csoportok lehetséges előfordulási száma n db változó esetén:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

Ez - láthatóan - a figyelemmel kisért változók számával exponenciálisan nő.

Ezen belül pl. egy adott k=3-as változócsoportban további p.q.r eshetőség van, ha a változók állapotának száma p, q és r.

Miután a gyakorlatban - egy-egy nagyobb felmérésnél - n értéke százas nagyságrendben mozog, a tulajdonságcsoportok előfordulási gyakoriságait csakis számitógépekkel érdemes kigyüjteni.

A szakembernek csak a válogatás főbb irányait kell megjelölni, az idő- és energia-pazarló rutinmunkát a gép végzi el. Természetesen a legkorszerübb gépi adatfeldolgozás sem nélkülözheti a gondolkodó embert ! Erre nemcsak az adatok betáplálását megelőző szelektálásnál, hanem az eredmények értékelésénél is szükség van. Fontos, hogy a nem vizsgált változók hatását eliminálju k !

Miután a biológiai vizsgálatoknál általában nincs módunk a jelenségeket tetszés szerint befolyásolni, irányitani, igy nem tudjuk lejátszani, modellezni ugy a folyamatokat, hogy az egyes tényezőket rendre kiküszöbölve, az egyedi hatásokat megállapithassuk. Ezért ugy kell válogatnunk az adatokból, hogy a nem vizsgált változók értékei állandóak legyenek (ill. egy elfogadhatóan szük intervallumon belül mozogjanak), hogy hatásuk ne érvényesüljön: ne fedhessék el a vizsgált kapcsolatot.

Pl. ha az X_1 (bemeneti) változónak az Y (kimeneti) ordinátára gyakorolt parciális hatását vizsgáljuk, ugy a többi X_2, $X_3 \ldots X_n$ változónak konstansnak kell lenniök. Ellenkező esetben ugyanis nem lehetünk biztosak abban, hogy Y megváltozásait csak X_1 megváltozásai okozzák - hiszen a "nem figyelt" X_2, X_3, $X_4, \ldots X_n$ tényezők is hatnak, akaratunktól függetlenek, objektivek.

Ha nem igy járunk el, ugy akár a valósággal homlokegyenest ellenkező eredmények is adódhatnak !

Igy a rendelkezésünkre álló adathalmazt valamennyi változó szerint, számtalan szempontból analizálhatjuk, - mely gépi uton könnyen és gyorsan elvégeztethető.

Természetesen az értékelés nagy körültekintést és megfelelő szakmai ismereteket kiván meg. Az egyes tulajdonságok, ismérvek együttes előfordulásának gyakoriságából nem lehet egyértelműen ok-okozati kapcsolatra következtetni. Lehet ritkán előforduló faktorok között is összefüggés - s megforditva is áll: a gyakran együttjáró tulajdonságok sem jelentenek mindig kauzalitást.

II. Mennyiségi ismérvek:

Amennyiben adataink kvantitativek, ugy a regresszió- III. faktoranalizis segitségével kereshetjük a vizsgált tényezők között fennálló kapcsolatokat.

A.) A regressziószámitás lényegében a mért változók valamennyi adatához legjobban illeszkedő függvény együtthatóit határozza meg. (Gauss-féle legkisebb négyzetek elve.) Lineáris összefüggés esetén az

$$Y = a_o + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

egyenlet együtthatóira teljesülnie kell az alábbinak:

$$\sum_{i=1}^{m} \left[Y_i - /a_o + a_1X_{1i} + a_2X_{2i} + \ldots + a_nX_{ni}/ \right]^2 = \varphi \stackrel{!}{=} \text{Minimum}$$

/m = esetszám/

Ebből:

$$\frac{\partial \varphi}{\partial a_o} \stackrel{!}{=} 0$$

$$\frac{\partial \varphi}{\partial a_1} \stackrel{!}{=} 0$$

$$\vdots \qquad \vdots$$

$$\frac{\partial \varphi}{\partial a_n} \stackrel{!}{=} 0$$

A többváltozós szélsőértékszámitás fenti egyenleteiből a keresett a_o, a_1, a_2,... a_n regressziós együtthatók adódnak.

Ezek, mint az egyes változók sulyozó együtthatói, megmutatják, hogy a kérdéses tényezők milyen mértékben és irányban befolyásolják a vizsgált folyamat eredményét, külön-külön.

A regressziós egyenletből tehát választ kapunk a jelenséget befolyásoló be- és kimenő változók, azaz az okok és okozatok közti közvetlen - a részeffektusokat nem tükröző - kapcsolat jellegére. Nem ismervén, illetve nem vizsgálván a folyamatot befolyásoló valamennyi tényező hatását, e kapcsolat természetesen csak mint tendencia juthat érvényre. Igy a regressziós összefüggés segitségével - egy jövőbeni mérés, illetve kisérlet eredményére csak egy megadott valószinüséghez rendelhető bizonytalansági közön belül következtethetünk. (Konfidencia-intervallum).

Ha a változók mérési tartományán kivül becslést alkalmazunk, ugy következtetéseinknek szakmailag is megalapozottnak kell lenniök. Ugyanis maga a statisztikai elemzés az extrapolációra nem jogosit fel.

Mig a regressziószámitás a változók közti kapcsolat jellegét, a korrelációszámitás annak szorosságát adja meg.

A többszörös korrelációs együttható az összes független változót egyként kezelve - méri azok és a függő változó közti kapcsolat mértékét.

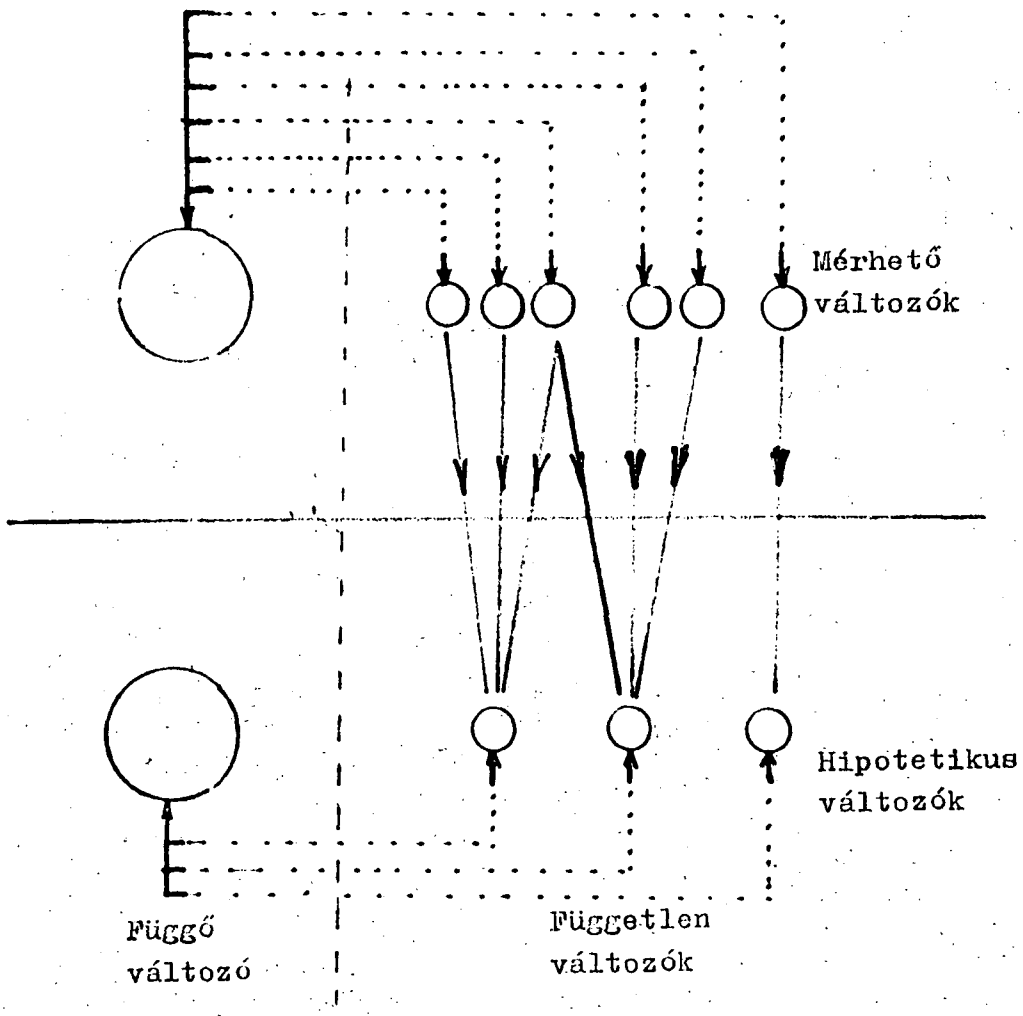
A parciális korrelációs együttható viszont az egyes független változók és a függő változók közti korrelációt jelzik. A többi független változót állandó értéken tartva, azok hatása nem érvényesül.

Ezzel szemben a totális korrelációs együtthatók a két kiszemelt változó közti közvetlen és közvetett hatásokat együttesen mérik. Ugyanis a kiválasztott két változó a többi változóval kapcsolatban van, - igy e mérőszámban az azokon keresztül érvényesülő hatásuk is kifejeződik.

B.) Faktoranalizis:

Szemben a regresszióanalizissel, mely a mérhető változók kapcsolatát elemezte, - a faktoranalizis már "mélyebb rétegekben" keresi az okozati összefüggéseket.

Sematikusan:



Ugyanis a vizsgált (mért) "független" változók valójában nem minden esetben függetlenek egymástól, mert ezek - illetve egy részük - számunkra még ismeretlen közös faktorok: un. "hipotetikus változók" függvényei. A mért változók ezeken a feltételezett közös faktorokon keresztül kerülnek egymással függőségbe.

Igy az egyes X_1, X_2, X_3,...X_n mért változók valamilyen K_1, K_2, K_3,...K_m közös faktorok lineáris kombinációjaként előállithatók:

$$X_1 = a_{11} K_1 + a_{12} K_2 + \ldots + a_{1m} K_m$$

$$X_2 = a_{21} K_1 + a_{22} K_2 + \ldots + a_{2m} K_m$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$X_n = a_{n1} K_1 + a_{n2} K_2 + \ldots + a_{nm} K_m$$

De ha a mért változók előállithatóak a közös faktorok felhasználásával, akkor megforditva is igaz: az egyes közös faktorok is kell, hogy előállithatóak legyenek a ténylegesen megfigyelt változók segitségével. Igy pl. a K_1-re irható:

$$K_1 = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n .$$

Tehát a közös faktorok - melyek jelenlétére a mért változók függőségéből következtettünk, - számszerüsithetőek, a mérési értékekből - az α_i együtthatók ismeretében megadhatóak.

Viszont számtalan esetben nehézséget okoz ezek biológiai értelmezése ! Ugyanis ha egy-egy hipotetikus változó a legkülönfélébb jellegü mért változók lineáris kombinációjaként adódik, - ugy nehéz annak biológiai tartalmat tulajdonitani.

Amennyiben igy egy ujabb kapcsolatot ismerünk fel, ugy hasznos lehet e módszer, egyébként pedig - nem mondván lényegesen többet a változók kapcsolatáról, mint a regresszióanalizis - megfontolandó, hogy indokolt-e ahelyett a faktoranalizis sokkalta bonyolultabb apparátusát belevonni vizsgálatainkba. Ugyanis ha a megfigyelt változók standardizált (z) értékére (amikoris azok átlaga zérus, szórásuk pedig egy) általánosan:

$$\underline{z} = \underline{\underline{A}} \cdot \underline{f} = \underline{\underline{A}}_k \cdot \underline{k} + \underline{\underline{A}}_u \cdot \underline{u}$$

ahol:

$\underline{z}$ = a standardizált változók oszlopvektora

$\underline{k}$ = a közös faktorok oszlopvektora

$\underline{u}$ = az un. egyedi faktorok oszlopvektora

$\underline{\underline{A}}$ = egyÜttható-mátrix

akkor ebből - részletesen kiirva - speciális alaku regressziós egyenletet kapunk.

Egy adott z_i változóra ez:

$$z_i = a_{i1}K_1 + a_{i2}K_2 + \ldots + a_{im}K_m + a_iU_i$$

Az ebben szereplő független (K) változók nem mérhetőek, igy az a_{ij} faktorsulyok sem határozhatóak meg a regressziós elemzés szokásos módszereivel - csak más, lényegesen bonyolultabb uton.

A részletek mellőzésével, az un. "alapvető faktorok módszere" (Principal Factor Solution) az alábbi sajátérték - sajátvektor problémára vezet:

$$\left(\underline{\underline{R}}_h - \lambda_1 \underline{\underline{E}}\right) \cdot \underline{a}_1 = \underline{\underline{0}}$$

azaz:

$$\underline{\underline{R}}_h \cdot \underline{a}_1 = \lambda_1 \cdot \underline{a}_1 \qquad \underline{\underline{R}}_h = \frac{1}{N} Z'Z' \ *$$

ahol:

$\underline{\underline{E}}$ = egységmátrix

N = az elemszám

$\underline{\underline{R}}_h$ = a megfigyelések redukált korrelációs mátrixa

(Megjegyzés: itt az un. egyedi faktoroktól eltekintettünk: a Z' erre utal.)

$\underline{a}_1$ tehát az $\underline{\underline{R}}_h$ redukált korrelációs mátrix egy sajátvektora.

Egyéb megszoritásokat is figyelembevéve az $\underline{a}_1$ vektort az $\underline{\underline{R}}_h$ legnagyobb saját-értékéhez tartozó saját-vektorok halmazából kell kiválasztanunk.

Ez éppen egy $\sqrt{\lambda_1}$ hosszuságu sajátvektor.

Az $\underline{\underline{A}}_k$ mátrix második $\underline{a}_2$ oszlopának meghatározása az $\underline{a}_1$ meghatározásához hasonlóképpen történik, azzal a különbséggel, hogy $\underline{\underline{R}}_h$ szerepét az

$$\underline{\underline{R}}_1 = \underline{\underline{R}}_h - \underline{a}_1 \underline{a}_1^*$$

un. első reziduális korrelációs mátrix veszi át. A továbbiakban is hasonlóképpen járunk el.

Az igy nyert $\underline{a}_1, \underline{a}_2, \ldots \underline{a}_m$ vektorok páronként ortogonálisak, és az eljárás során adódó λ sajátértékekre fennáll:

$$\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_m$$

Az utolsó $\underline{\underline{R}}_m$ reziduális mátrix értéke zérus, igy a fenti eljárás teljes mértékben reprodukálja a redukált korrelációs mátrixot.

A vázolt módszerrel tehát az a_{ij} faktorsulyok - s igy - átrendezés után az α_i keresett együtthatók is adódnak. Azaz a mért változók segitségével ez utóbbiak ismeretében a K_i közös faktorok meghatározhatók.

Hogy az ismertetett módszerek közül mikor melyiket célszerübb használni, azt esetenként kell eldönteni. Intézetünkben (Pécsi Orvostudományi Egyetem Egészségügyi Szervezéstani Intézet Számitástechnikai Csoport) az elmondottak figyelembevételével dolgoztuk ki a számitógépes adatfeldolgozó programokat.

E programokkal vizsgáljuk szociális körülmények, társadalmi adottságok és a különféle betegségek kapcsolatát.

Segitségükkel elemezzük a koraszülést befolyásoló tényezőket is. Ugyanakkor lehetővé teszik a klinikákon, kórházakban felgyülemlett hatalmas adathalmaz korszerü, gyors feldolgozását, áttekintését: kezelési eljárások, gyógyszerhatások és mellékhatások, laboratóriumi eredmények, diagnózisok stb. közti összefüggések felismerését.

I r o d a l o m

1.) Ezekiel, M. - Fox, K.: Korreláció és regresszió analizis (lineáris és nem lineáris módszerek)
Közgazdasági és Jogi Kiadó Budapest, 1970.

2.) Prékopa András: Valószinüségelmélet. (Müszaki alkalmazásokkal.)
Müszaki Könyvkiadó Budapest, 1972.

3.) Vincze István: Matematikai statisztika. (Ipari alkalmazásokkal.)
Müszaki Könyvkiadó Budapest, 1968.

4.) Vita László: A faktoranalizis közgazdasági alkalmazásának lehetőségeiről.
Szigma 1970. III. évf. 2.sz. 127-152 old.